

Action Recognition in Videos: from Motion Capture Labs to the Web

Ana Paula Brandão Lopes^{1,2}, Eduardo Alves do Valle Jr.³, Jussara Marques
de Almeida¹, Arnaldo Albuquerque de Araújo¹

¹*Depart. of Computer Science – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte (MG), Brazil*

²*Depart. of Exact and Tech. Sciences – Universidade Estadual de Santa Cruz (UESC)
Ilhéus (BA), Brazil*

³*Institute of Computation – Universidade Estadual de Campinas (UNICAMP)
Campinas (SP), Brazil*

Abstract

This paper presents a survey of human action recognition approaches based on visual data recorded from a single video camera. We propose an organizing framework which puts in evidence the evolution of the area, with techniques moving from heavily constrained motion capture scenarios towards more challenging, realistic, “*in the wild*” videos. The proposed organization is based on the representation used as input for the recognition task, emphasizing the hypothesis assumed and thus, the constraints imposed on the type of video that each technique is able to address. Expliciting the hypothesis and constraints makes the framework particularly useful to select a method, given an application. Another advantage of the proposed organization is that it allows categorizing newest approaches seamlessly with traditional ones, while providing an insightful perspective of the evolution of the action recognition task up to now. That perspective is the basis for the discussion in the end of the paper, where we also present the main open issues in the area.

Keywords: Action Recognition, Survey, Video Analysis

Email addresses: paula@dcc.ufmg.br (Ana Paula Brandão Lopes^{1,2}),
mail@eduardovalle.com (Eduardo Alves do Valle Jr.³), jussara@dcc.ufmg.br (Jussara
Marques de Almeida¹), arnaldo@dcc.ufmg.br (Arnaldo Albuquerque de Araújo¹)

1. Introduction

This paper presents the state-of-the-art in action recognition for videos based on visual data recorded from a single camera. It shows how the approaches have evolved from the analysis of videos produced in heavily constrained motion capture environments towards recent attempts of automatically understanding realistic or “*in the wild*” videos. Our goal is not to provide an exhaustive survey, but rather an elucidative overview of the main ideas in the area, its historical evolution and its current trends.

A number of existing review papers more or less related to the task of action recognition in videos provide different perspectives on the field. Put together, they also provide the perspective of its evolution (Section 2). However, most of them fail to cover the most recent achievements of the area. The few exceptions inherit the traditional organization and taxonomy of older surveys, which are unable to characterize the current corpus of methods adequately.

In this survey, a new organizing scheme including both old and recent approaches is proposed in Section 3. For the sake of completeness, an overview of older approaches is presented in Section 4, while recent approaches are covered in deeper detail in Section 5. A summary followed by a discussion on current research trends is provided in Section 7. Finally, concluding remarks are presented in Section 8.

1.1. Why Should We Recognize Actions in Videos?

In recent years, Internet users witnessed the emergence of a great amount of multimedia content in the Web. Initially, such kind of content was generated by professional or semi-professional individuals or enterprises, in a typical broadcasting scheme. However, in a second wave, the users themselves started creating and publishing their own multimedia materials. This motion towards user-generated content was motivated by a number of factors, mainly the drop in the cost of devices such as cameras and microphones and the spread of high-bandwidth connections, as well as the emergence of Web 2.0 applications, including online social networks. This new scenario led to an overwhelming increase in the amount of multimedia content available, which by its turn brought up the limitations of traditional Web tools in dealing with non-textual data.

To make multimedia data effectively available, the high-level indexing aimed at meaningful, semantically-oriented retrieval is a critical goal. While

for text, the words themselves convey quite directly its semantics, in the case of visual information, the connection between low-level encoding (i.e., pixels) and semantic meaning is far from immediate. Indeed, it is an open research issue, commonly referred to as the *semantic gap* [1].

The current state-of-the-art in terms of systems for image and video retrieval is described in [2]. Such systems are composed by several single individual concept detectors which are applied independently to every item in the database. The estimated probabilities of occurrence then compose the feature vectors for the search engine. Systems like the one just described have the advantage of enabling textual search, as opposed to query-by-example¹ or query-by-sketch² approaches, which are not always intuitive for users accustomed to commercial search engines.

One key issue in such approach is which concepts should be considered. This issue was addressed in the Large-Scale Concept Ontology for Multimedia (LSCOM) workshop [3], which defined a lexicon containing around 1000 concepts, from which 449 have been annotated in 80 hours of video coming from the TREC Video Retrieval Evaluation (TRECVID) 2005 database [4].

Later on, experiments performed by [5] showed that annotations for some concepts defined in LSCOM varied significantly whether the annotators watched video sequences or looked at keyframes. Those results indicate that the dynamic nature of video information plays an important role in the recognition of some concepts. Also, they suggested that such dynamic semantics cannot adequately be captured by the direct application of techniques aimed at still images.

The 24 activity/event LSCOM concepts which had their annotations changed after the experiments described in [5] are listed in Table 1. From that table, it is possible to see that all those concepts, either directly or indirectly, are related to actions performed by human beings.

Additional Applications

Although the improvement of high-level video indexing and retrieval is an important motivation for action recognition research, it is worth mentioning several additional applications.

¹In query-by-example systems, the users choose an image as a query and the system returns those ones considered most similar to it.

²In query-by-sketch systems, the users need to draw a rough sketch of the image they want to find.

Table 1: LSCOM concepts that were found highly dependant on motion. Users frequently re-annotated them when switching to viewing video segments instead of keyframes [5].

Airplane Crash	Greeting
Airplane Flying	Handshaking
Airplane Landing	Helicopter Hovering
Airplane Takeoff	People Crying
Car Crash	People Marching
Cheering	Riot
Dancing	Running
Demonstration Or Protest	Shooting
Election Campaign Debate	Singing
Election Campaign Greeting	Street Battle
Exiting Car	Throwing
Fighter Combat	Walking

A great amount of work has been done around the idea of building “smart” video surveillance systems, which would be able to detect suspicious behavior automatically. In [6], for instance, a framework to aid the search for specific events in recorded surveillance video is proposed. In addition, recognition of people by their gait has been studied as alternative biometrics [7]. A review focused on visual surveillance systems is presented in [8].

The analysis of sport videos is another important application. In [9], for example, the classification of video segments between play and break intervals is suggested to summarize the video, by taking out the breaks. Soccer games are also analyzed by [10], in which text and the players’ trajectories are used to build a system aimed at helping coaches in tactical analysis. Six actions of a cricket umpire are analyzed in [11] – by a technique using an appearance-based method similar to eigenspaces (commonly used in face recognition) – whereas the usage of local motion analysis to identify different swimming

styles is proposed in [12].

Hand gesture recognition can be useful for a number of applications. In [13], it is applied to identify which segments in lecture videos are worth transmitting in less compressed formats. The underlying assumption of that work is that specific actions can indicate the importance of each sequence, therefore guiding a semantically-oriented compression. Automatic recognition of sign language symbols is explored in [14] and [15], for example. In [16], the recognition of hand manipulations of objects recorded by a camera attached to a person body are suggested as a means of interaction with a virtual reality system.

Human-Computer Interaction (HCI) systems can also benefit from the ability of recognizing generic actions, as it can be seen in the pioneer work of [17]. They present *KidsRoom*, an environment able to interpret and react to specific actions of a group of children in a closed space. In a similar application, [18] proposed a system called *smart classroom*, where the actions performed by a teacher are recognized to allow automatic camera motion and a virtual mouse. Facial actions have been recently explored either as a tool to enhance HCI – as in [19] – or to analyze the affective behavior of psychiatric patients [20].

A specific instance of the general content-based retrieval idea is what is called Smart Fast-Forward (SFF), as proposed by [21], in which the query video segment is compared against other segments in the same video in order to find similar actions taking place in different intervals.

Action recognition is an important issue also in robotics, in which the interpretation of human actions can be used either for reaction to the recognized action (i.e., control) or for learning and imitation [22].

Finally, in the medical area, human motion analysis can aid diagnosis of motor problems by comparing patient motion to normality patterns, as in [23], for example. Another possible medical application is to provide remote assistance to elderly people, as suggested in [24]. Similar medical applications also motivate the work of [25].

2. Related Surveys

An extensive survey on earlier studies about motion-based recognition was first presented in [26]. For the authors of that work, the first step in motion-based recognition is the extraction of motion information from a sequence of images, which can be done by optical flow or motion correspondence. Motion

correspondence is established by tracking specific points of interest through frames, and generating motion trajectories, which can be parameterized in several ways. Instead of computing motion information from the entire image or from specific points, region-based motion features can also be extracted. Explicit human body models are used to guide the tracking step.

The survey presented in [27] is devoted to human motion analysis which, for them, comprises the following overall steps: a) segmentation; b) joint detection; and c) identification and recovery of 3D structures from 2D projections. The authors characterize body structure analysis as either model-based or non-model-based, depending on whether or not an *a priori* shape model is used. The *a priori* models considered can be stick figures, contours or spatio-temporal volumes. The proposals reviewed in [27] are also split into two broad categories – more related to action modeling and recognition steps (see Figure 1 in Section 3) – which are: *space-state models*, in which each static posture is considered as a state and state transitions occur with certain probabilities; and *template matching models*, where a template is computed for each action, and then a nearest-neighbor classification scheme is applied to recognize similar actions.

While the authors in [27] focus on approaches based on models of the human body, a survey specifically focused on models for recognition of hand gestures is presented in [28].

By their turn, [29] reviews approaches modeling either the whole body or the hand. Selected papers are organized into three categories: *2D approaches without explicit shape models*, *2D approaches with explicit shape models* and *3D approaches*. This survey also provides (in its Table 1) a comprehensive list of applications envisioned at the time, organized into five groups: virtual reality, smart surveillance systems, advanced user interfaces, motion analysis and model-based coding. It is worth noticing that most of the prospective applications suggested by [29] still remain as unsolved challenges.

In [30], the recognition of human action is described as comprising the following – more general, in comparison to [27] – steps: a) extraction of relevant visual information; b) representation of that information in a suitable form; c) interpretation. The specific modeling of human body or body parts is not seen by the author as an essential step for human action recognition. In contrast, tracking and trajectory computation are considered the primary subtasks. Therefore, this survey is focused specifically on trajectory-based techniques.

For [31], human motion analysis comprises the following steps: a) mo-

tion segmentation; b) object classification of segmented moving regions – which can be shape or motion-based; c) tracking of identified objects along consecutive frames; and d) recognition of motion patterns, providing what they call *behavior understanding*. As in [27], action modeling approaches are distinguished between *template-based* and *space-states based*.

The authors of [32] present a survey from the perspective of the generative learning algorithms applied to any of the various processing steps of action understanding systems. In that paper, such systems are categorized generically into *explicit models* and *exemplar-based models*.

In [33], the focus falls again onto approaches relying on human-body or body parts. That survey is mainly motivated by biometrics applications, and the paper is composed of two main parts: in the first, the author provides a detailed survey on tracking techniques applied to heads, hands or the whole body. In the second, techniques for analyzing different models for those tracked elements are reviewed.

The work of [34] expands and updates the earlier review presented in [27], by including not only actions, but also interactions. The surveyed approaches are distinguished by the level of detail in which the moving objects are described. *Coarse level approaches* are those in which people are considered as bounding boxes or ellipses. Then, motion patterns are used to model the actions. In approaches lying in the *intermediate level* of detail, people are represented by large body parts or silhouettes. Finally, *detailed models* can be built on the entire body or on specific parts, such as hands in the case of gesture recognition tasks.

Action modeling approaches are also distinguished based on two different aspects: the first differentiates *direct recognition* from reconstruction of *body models* before recognition. The second aspect distinguishes approaches by their *static* or *dynamic nature*, if the recognition is performed on a frame-by-frame basis or taking the entire sequence as the basic unit analysis. High-level recognition schemes – similar to those which [31] call behavior understanding – are also discussed, most of them relying on manually constructed semantic models of the world.

In [8] an extensive review on papers related to surveillance systems is provided. The authors consider a visual surveillance system comprising of the following steps: a) motion detection, which includes object modeling, segmentation and classification; b) tracking of moving objects; and c) behavior understanding. For some applications, an additional step of natural language description can be added. The ability to identify people at a distance

(gait-based recognition) can also be introduced.

In another review focused on trajectory-based approaches, [35] define what it is called *activity inference*, comprised, in their view, of three steps: a) low-level video processing; b) trajectory modeling; c) similarity computation. In this scheme, low-level processing is aimed at computing trajectories for selected objects. The trajectories for each action can be modeled by varied techniques and for each model a similarity measure needs to be established.

The survey of [36] offers an overview of human motion analysis in general, with a section devoted to action recognition. They suggest that action recognition approaches can be broadly separated between the ones that explicitly consider human presence in the scene and the ones that do not. The recognition section of that paper is structured around three different kinds of tasks: *scene interpretation*, without identifying particular objects; *holistic recognition*, using the human body or body parts, to recognize both the subjects and the actions performed by them; *action primitives and “grammars”*, in which motor primitives are used for representation or control. The primitives in the latter task can be used to create an action hierarchy that gives a semantic description of the scene. However, in most of such approaches motion primitives are usually taken as already available.

The review of [22] is focused on robotics applications, more specifically for learning and imitation. They distinguish approaches based on: *scene interpretation*, in which the moving objects are not “identified”, but have only their overall motions analyzed; the *body as a whole*; *body parts* and *grammars*.

The review presented in [37] deals specifically with pose estimation, assumed as a needed step for action recognition.

In the recent short review presented in [38], which is explicitly focused on papers from 2001 to 2008, a hierarchical terminology composed of *action primitives*, *actions* and *activities* is adopted. This survey categorizes different proposals according to the Machine Learning (ML) techniques applied, regardless of the underlying representation. In other words, it is focused on the modeling step (Figure 1).

The major branches presented in [39] differentiate between approaches aimed at *actions* and those aimed at *activities* recognition. In their case, similarly to [36], actions are defined as simple motion patterns executed by a unique human, while activities are more complex patterns, normally involving more than one person. The following four major steps for action recognition are identified: a) collecting input video; b) extracting low-level

features; c) extracting mid-level action descriptions; and d) high-level semantic interpretations.

The low-level features considered in that survey are *optical flow*, *point trajectories*, *blobs and shapes separated from the background* and *filter responses*. According to them, actions can be described at mid-level by *non-parametric*, *volumetric* and *parametric* models. Actions and activities can be modeled either by *graphical models*, *syntactic grammars-like approaches* or *knowledge/logic based approaches*.

The work of [40] reviews motion recognition approaches in the context of Content-Based Video Retrieval (CBVR). Two major approaches are identified. In *trajectory-to-trajectory approaches*, motion trajectories are extracted and compared for recognition; the category of approaches that take into account the internal structure of the object over time are denominated *sequence-to-sequence approaches*.

In [41], only papers aimed at recognizing full body actions are taken into account. Image representations are separated into three large groups: *global*, when a specific Region of Interest (ROI) is described globally, *local*, based either on interest points and densely sampled ones, and *application specific*. In his view, action classification can be performed either by *direct classification*, using the information coming from all the frames in the sequence together and *temporal state-space models*, in which action sequences are broken in smaller steps.

3. Categorizing Different Approaches for Action Recognition

Regardless of the application envisioned, the process of recognizing human actions from videos can be seen as comprising the three major steps, as depicted in Figure 1.

- (a) ***Representation Extraction***: this step starts with the extraction of low-level features from the videos, like color, texture and optical flow, for example. Those features are usually fed to a somewhat complex processing chain until a suitable (i.e., compact and descriptive) representation is achieved. It is worth noticing that, unlike [39], for instance, the output of this step is the final video representation which is used as the input to the action modeling step (below), regardless of the abstraction level. This generic definition is then applied to the finer-grained hierarchical structure proposed later in this section.

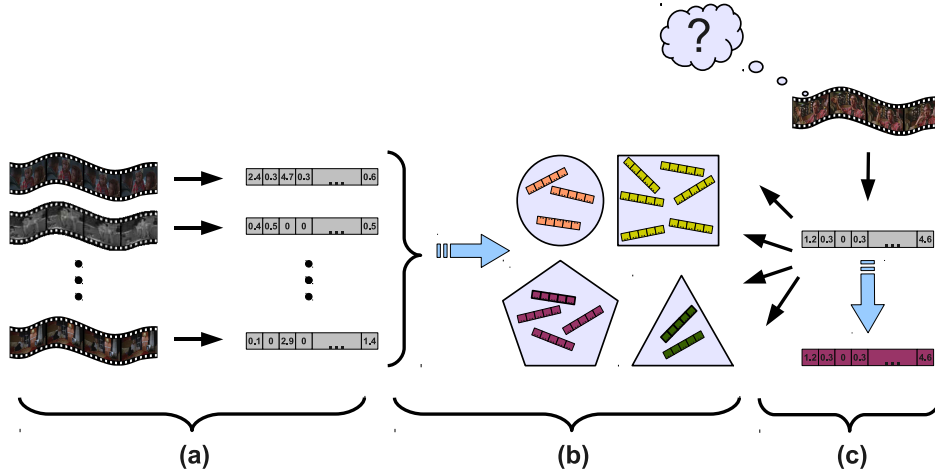


Figure 1: Overview of the processing steps needed for action recognition in videos: (a) *representation extraction* step, (b) *action modeling* step, and (c) *action recognition* step (picture best viewed in color).

- (b) **Action Modeling**: in this step, the representations built in the previous step are mapped into different action categories. The *spectrum* of modeling alternatives goes from the selection of a small number of action templates aimed at direct comparison to sophisticated modeling schemes involving ML techniques.
- (c) **Action Recognition**: this last step takes place when unlabeled (query) videos are analyzed against the previously built action models, so that those videos can be associated with one of the possible action categories (i.e., classified).

As expected, those three steps are tightly interconnected. Some representation choices are more suitable for – or are even designed specifically to go with – certain kinds of techniques for action modeling. In the same way, the selected action modeling technique will determine – in some cases, in a unique way – how the classification is going to be performed.

The structure proposed in this section for organizing different approaches for action recognition is based on the representation step depicted in Figure 1. Such choice is justified by the fact that the process of extracting a specific representation for videos is directly related to a number of assumptions about the scene content. Such assumptions, by their turn, impose specific constraints on the types of videos that each recognition approach is

able to cope with. Hence, an organization of different approaches which is built based on the selected representation provides a better ground to understand the strengths and limitations of each category of approaches, making it easier to: a) select appropriate approaches for specific applications, and b) distinguish which approaches are truly comparable among them. Finally, the selected *criteria* based on the underlying representation allow for sensibly unraveling unrelated approaches that end up mixed together under other categorization schemes.

Regardless of all the existing surveys discussed in Section 2, authors follow no standard categorization structure while referring to previous papers that are related to their proposed approaches. For instance, in [42], authors propose a broad categorization between *object centric* and *statistical* approaches, while in [43], authors distinguish among approaches based on *3D tracking of different points* of human bodies, *accurate background subtraction*, *motion descriptors on regions of actions* and *learning of actions models*. In [44], human action approaches are categorized into those based on *tracking*, *flow*, *spatio-temporal shapes* and *interest points*. In [45], different approaches are categorized as *model-based*, *spatio-temporal template-based* and *bags-of-visual-features-based*. In the work of [46], previous papers are coarsely classified according to their specific goals, distinguishing among approaches dealing with *unusual event detection*, *action classification* and *event recognition*.

The categorization scheme we propose in this paper is depicted in Figure 2. In a coarse level, the different approaches are split into two large groups, which are nearly equivalent to the object centric *versus* statistical categorization proposed by [42]. It can also be considered a generalization of the categorization of [36] into approaches that either consider the presence of humans or not. In fact, the framework proposed in this paper is a refinement of the model-based *versus* model-free categorization presented in [47], although the terms *model-based* and *model-free* are abandoned in order to avoid confusion between action modeling and object modeling, the former being always present (Figure 1). The proposed scheme stresses the distinction between approaches that explicitly assume the presence of moving objects under specific conditions – like, for example, homogeneous background – from those in which such explicit assumption is not found.

As it will be seen, there is a non-negligible correlation between the proposed taxonomy and the temporal evolution of approaches: more recent approaches tend to rely on less constrained assumptions and therefore, more

general hypothesis.

The two initial categories are further refined into subcategories organized according to the underlying representation. The vast majority of proposed solutions to human action recognition to date lies in the first broad group of approaches depicted in Figure 2. In other words, the video representation used by them explicitly assumes that one or more moving objects appear in the scene, typically under a number of specific conditions, like stable backgrounds and fixed scales, for example.

The basic idea behind those approaches is that it is possible to infer the actions being performed by studying the structure and/or the dynamics of the moving objects in the scene (or their parts). Moving objects of interest can be the human body, some body parts or other objects related to the application domain, like airplanes and automobiles, for example. Unlabeled moving regions can also be considered. In order to be able to analyze the moving objects, they need to be detected (and often, also tracked) before any further processing. Once the object has been detected/tracked, it can be either a) adjusted to some pre-defined model of the object, characterized by a number of parameters (parameterized object models), or b) characterized by global descriptors computed on their segmented area (implicit object models). Approaches relying on the presence of specific moving objects in the scene are further discussed in Section 4.

More recently, a number of approaches that do not explicitly rely on the presence of any specific object in the scene have been proposed. They are based on global statistical computations over different kinds of representations, within distinct abstraction levels: low-level features, mid-level interest points and high-level concepts. Approaches based on global statistics are represented by the lowest large rectangle in Figure 2 (which starts the branch in turquoise). As such approaches have only recently been proposed, previous surveys, discussed in Section 2, do not cover them in much detail. We provide a more comprehensive discussion of those approaches in Section 5.

Finally, hybrid approaches, which mix ideas from both those ones based on models of pre-detected objects and those ones based global statistics can also be found.

Tables 2, 3 and 4 summarize the approaches for action recognition discussed throughout this paper, pointing out the main assumptions for each category, and citing related papers that are going to be discussed in the following sections.

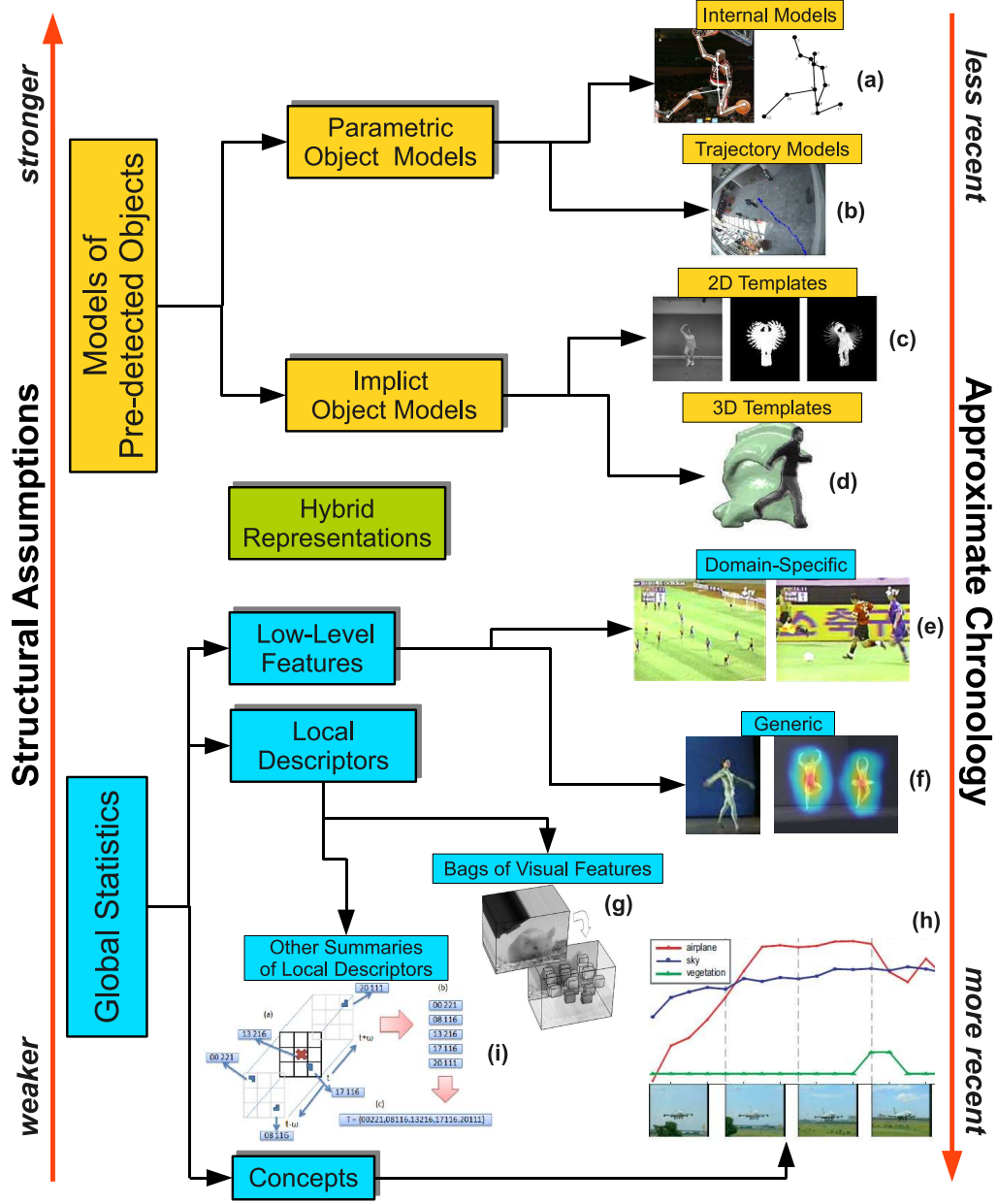


Figure 2: Categorization framework used along this survey for organizing the approaches for human action recognition found in the literature. It is based on the underlying video representations. Image references: (a) [48], (b) [49], (c) [50], (d) [51], (e) [9], (f) [52], (g) [53], (h) [54], (i) [55] (picture best viewed in color).

Table 2: Summary of Action Recognition Approaches Based on Object Model Representations (table best viewed in color).

Basic Representation – Object Models			
<i>Main Assumption: Actions can be derived from a specific model of the related objects.</i> [Related Papers by Subcategory (below)]			
Parametric Models		Internal Models	
<i>Objects related to actions obey a predefined model.</i>		<i>Global representations of objects internal areas implicitly define the object model.</i>	
Internal Models	Trajectories	2D Descriptors	3D Descriptors
<i>A predefined object model must describe objects internal states.</i>	<i>The relevant information is not in the objects internal state, but in their positions over time.</i>	<i>Objects appearance information is enough for action recognition OR motion can be aggregated in 2D representations.</i>	<i>Changes in appearance over time are also relevant for action recognition.</i>
[48] [56] [57] [58] [59] [60] [61] [62]	[49] [63] [64] [65] [66]	[50] [67] [68] [69] [70] [71]	[51] [72] [73] [74] [75] [44] [43]

4. Approaches Based on Models of the Moving Objects

In this section, approaches which do assume that specific objects are in the scene under constrained conditions are discussed according to the structure proposed in upper part of Figure 2. This category of approaches relies on models for those objects assumed to be performing the actions, models that can be either explicit (parametric) or implicit.

As indicated in Figure 2, approaches based on object models are those that appeared first in the literature, out of traditional motion capture research. Some recent and/or classical papers of each branch are cited as needed throughout the following text, but the list presented is not meant to be exhaustive, since those approaches are broadly covered in previous surveys (Section 2). Therefore, this section is mainly aimed at providing a high-level overview of that category of approaches, detailing it enough to give the reader some perspective on the evolution of the area towards less constrained techniques.

Table 3: Summary of Action Recognition Approaches Based on Global Statistical Representations (table best viewed in color).

Basic Representation – Statistics				
<i>Main Assumption: Global statistics capture relevant information for action recognition.</i>				
[Related Papers by Subcategory (below)]				
Low-level Features		Local Descriptors		Concepts
<i>Low-level features can indicate actions being performed.</i>		<i>Mid-level local descriptors are better suited to capture relevant information.</i>		<i>Concepts occurrences can indicate the actions being performed.</i>
Domain-Oriented	Generic	BoVF	Other	
<i>Domain information must guide the choice of relevant low-level features.</i>	<i>Generic low-level features are able to capture relevant information.</i>	<i>A histogram of quantized local descriptors can be associated with actions.</i>	<i>Other ways than BoVF can better capture relevant information from local descriptors.</i>	
[9] [76] [77] [78]	[6] [79] [80]	[89] [90] [91]	[99] [100]	[42] [106] [107] [46]
	[81] [21] [52]	[53] [92] [93]	[101] [102]	
	[82] [83] [84]	[94] [95] [45]	[55] [103]	
	[85] [86] [87]	[96] [97] [49]	[104] [105]	
	[88]	[98]	[25]	

4.1. Using Parametric Object Models

The approaches based on parametric object models are the ones more directly related to motion capture techniques. In such approaches, moving objects (e.g., the human body, the hand, cars in a parking area) are assumed to follow a specific model and visual data is matched against that model to infer its parameters. In the action modeling step, different sets of values for the model parameters are then associated with different actions.

Approaches based on parametric models can be split into two major subgroups: those which parameterize the object internal structure, and those which ignore such internal structure and instead, parameterize only objects trajectories.

Table 4: Summary of Hybrid Action Recognition Approaches (table best viewed in color).

Basic Representation – Hybrid						
<i>Main Assumption: combinations of different approaches produce enhanced recognizers.</i>						
[Related Papers (below)]						
[108]	[109]	[110]	[111]	[112]	[113]	

Models of the Objects Internal Structure

These approaches, which are among the earliest techniques developed for action recognition, are mostly derived from motion capture techniques. Thus, they start by defining a detailed model of the internal structure of a pre-defined moving object and then adjusting the visual data to that model. The most commonly modeled objects are the entire human body as well as body-parts, such as hand models aimed at gesture recognition, for example. A classical example of human-body model (the so-called stick model) can be seen in Figure 2(a), from [48].

Strictly speaking, a detailed explicit model would require a great amount of 3D data, like in [48], thus leading to a high computational complexity. In addition, in most real-world scenarios, 3D data is simply not available. Hence, a number of approaches avoid that requirement by using simplified models. In [56, 57], for example, a simplified stick model is obtained from silhouettes. Poses are modeled based on a few points from the body in [58]. In [59] the human body is considered as a cooperative team of agents where each team member is a limb of the body. In [60] and [61] constellation models are employed to describe human poses. In [62], body parts are tracked using mixture particle filters and clustering the particles locally.

Trajectory Models

In trajectory-based approaches, the global motion of the objects is considered the only relevant information for action recognition. In other words, the internal state of the moving objects is ignored and such objects are represented mainly by their position tracked over time. Actions are then modeled by trajectories parameters, which, in turn, can come from a number of different trajectory models. Trajectory-based approaches are very common in surveillance scenarios – such as the one depicted in Figure 2(b), from [49]. A number of surveys specifically devoted to them were already discussed in

Section 2.

In [63], for example, it is argued that activities can be modeled by any representative shape associated with the activity to be modeled, and as an example case, the shape of the trajectories of a set of points associated with the moving object is analyzed. Instead, a large number of proposals analyze the trajectory of a unique point, as in [64], which is aimed at identifying common office activities by the analysis of hand trajectories. In [65] and [66], the focus is on detection of abnormal events in crowded scenes. In such scenes, tracking the objects of interest is particularly challenging. To overcome this, global motion fields are analyzed in order to discover *super-tracks*, which are intended to capture predominant motion patterns that are then used to model events.

4.2. Using Implicit Object Models

In this class of approaches, the area around the moving object — like a silhouette or a bounding box, for example — is detected and submitted to some kind of global description. This line of work assumes that explicit details of the object structures are not necessary for action recognition. Rather, the global features of a ROI defined around the object implicitly capture its model, at a lower cost. The lower computational effort offered by this basic idea gave rise to a variety of similar approaches, which can be distinguished between those using 2D templates and those exploiting spatio-temporal 3D templates.

Implicit Object Models Based on 2D Templates

A landmark paper using implicit object modeling as the basic representation is [50]. In this paper, two different 2D templates computed from extracted silhouettes is proposed: (a) Motion Energy Image (MEI), which is a binary image indicating where the motion occurred during the sequence; (b) Motion History Image (MHI), which is a gray level image where brighter pixels indicate the recency of motion (in other words, the brighter the pixel, the more recent the motion occurred there). Both MEI and MHI images are described by seven Hu moments, which are meant to carry a coarse shape description which is invariant to scale and translation (Figure 2(c)). The MEI/MHI representation proposed by [50] became the basis of a great number of extensions and variations, mostly applied to scenarios with relatively stable backgrounds (like the work of [69] aimed at surveillance applications).

Another classical approach using 2D templates as primary video representations appears in [67], which addresses the problem of recognizing human actions from medium resolution videos. This approach relies on the detection and stabilization of a bounding box containing the human figure. The description of such boxes is based on the optical flow projected into motion channels, which are blurred with a gaussian filter to reduce the sensitivity to noise which is typical of optical flow estimations. Such motion descriptors are later used by several other authors (see [86] and [74], for instance).

In [70], the internal part of previously detected motion regions are described by Bag of Visual Features (BoVF), a statistical representation based on interest points that is further detailed in Section 5.2.1.

Along with the global spatial descriptors, the information encoded in the sequential nature of video is explicitly taken into account by some authors relying on 2D templates. In [68], a bounding box centered at the moving body is described based on radial Histograms of Gradients (HoG). Such histograms are then clustered to create a codebook of poses, based on which each video is described by two alternative representations, namely: bag-of-poses and sequence-of-poses. A similar approach – using Non-negative Matrix Factorization (NMF) to build a bag-of-poses representation – is presented in [71].

Implicit Object Models Based on Spatio-temporal 3D Templates

In this category, actions are represented as 3D volumes in space-time. Such spatio-temporal volumes are created by aligning and stacking 2D information (e.g., silhouettes, contours, bounding boxes). The exploration of space-time volumes built on silhouettes for action recognition was first proposed in [51]. In their proposal, the properties of the Poisson equation are used to create a representation in which the values reflect the relative position of each internal position in the volume (Figure 2(d)).

The principle behind such approaches is that spatio-temporal volumes contain both static and dynamic information and are thus better suitable as representations for action recognition. Similarly to what happened with 2D template-based approaches, the initial idea of [51] is further explored in a number of subsequent papers, which describe the spatio-temporal volume using different techniques. For instance, [72], characterize the spatio-temporal volumes by their 3D geometric moments. In [73], it is proposed an algebraic technique for characterizing the topology of those volumes. Finally, the representation used by [74] can be considered an extension of the work of [67]

to a 3D spatio-temporal volume.

In [75] and [44], the authors explore space-time volumes at a smaller scale. Videos are over-segmented in space-time, creating micro-volumes, which are described based on optical flow information. They are then compared against manually segmented action templates, by a shape-matching technique adjusted to deal with over-segmentation.

In order to distinguish between drinking and smoking actions, [43] use densely sampled HoG and Histograms of optical Flow (HoF) 3D descriptors computed over manually cropped regions around people’s faces, used as input to a cascade of weak classifiers learned by AdaBoost.

5. Approaches Based on Global Statistics

Approaches which rely on the detection of moving objects share the drawback of depending on computer vision tasks – such as background segmentation and tracking – which are themselves open research issues. The lack of general solutions to those tasks leads to an excessive number of assumptions about what is in the scene, which ultimately makes such approaches applicable only to very constrained scenarios.

To cope with more realistic and unconstrained settings, different approaches make no assumption on the presence of any specific object in the scene, thus making object detection unnecessary. Instead, those approaches compute global statistics on different data. Statistics on *low-level features* (such as color, texture and optical flow, for example) can be computed either as generic descriptors or guided by specific information about the application domain, as further discussed in Section 5.1. Mid-level *Local descriptors* built on low-level data around selected points gave rise to an important branch of approaches based on Bag of Visual Features (BoVF), which are essentially histograms of quantized local descriptors. BoVF-based approaches are thoroughly discussed in Section 5.2.1. Although BoVF-based approaches dominated the scenario of statistical approaches in recent years, alternative proposals to gather information coming from local descriptors can be found and are discussed in Section 5.2.2. A third research line exploits the probabilities of high-level semantic concepts (e.g., sky, airplane, people) appearing in a video to infer the action taking place in it. These approaches are detailed in Section 5.3

5.1. Using Statistics of Low-level Features

In this category of action recognition approaches, low-level features of the video are statistically summarized and such summary is used as the video representation. Since the direct usage of low-level features is prone to suffer more intensely the effects of the semantic gap, some authors use previous knowledge about the application domain to guide the choice of features. Generic features without specific links to the application domain have also been exploited, although such approaches tend to be focused on a handful of constrained settings.

5.1.1. Low-level Statistics Guided by Domain Knowledge

A combination of low-level features and some previous domain knowledge is common in scenarios where the possible backgrounds are limited in number and have distinct global appearance. In professional sport videos, for instance, camera effects are commonly related to specific events. This fact is used in [9], in which dominant color ratio and motion intensity are computed to segment soccer videos between play and break intervals.

Another application for approaches based on low-level statistics is explored in [76], which use global motion information to identify generic, coarse-grained events in news videos, like *anchor*, *reporting*, *reportage* and *graphics*.

Many approaches based on low-level statistics appear as part of multi-modal frameworks (see, e.g., [77]), in which audio-visual features are mixed with high-level information to detect events in tennis games. The full exploration of multimodal frameworks is outside the scope of this paper, although we point the reader to [78] for a related survey.

5.1.2. Generic Low-level Statistics

A variety of approaches for computing generic global statistics based on low-level features, which do not rely on domain information, have been proposed in the recent literature. In most cases, they relate to constrained applications.

In [6], a framework aimed at searching for suspicious actions represents candidate video segments by histograms of intensity gradients both in spatial and temporal dimensions, over four different temporal scales. In [79], the typical dynamics of a surveilled environment is captured by statistical analysis of a 2D field containing the maximum activity of pixels. From that field, a model for normal behavior is produced, allowing comparisons with other videos so as to detect abnormal activities. Surveillance scenarios are

also the focus of [80], which propose a dynamic texture descriptor based on local binary patterns extracted from the three orthogonal planes formed by the spatial and temporal axes. The videos are represented by sequences of such descriptors computed over subsequent spatio-temporal subvolumes.

In [81] and [21], a generic approach for what is called Smart Fast-Forward (SFF) is presented. Their approach is based on the absolute values of normalized gradients computed over all space-time points, extracted in a temporal pyramid, to cope with different temporal scales. Points with gradients below a threshold are ignored in order to save time, and the remaining ones are described by the gradient components in x , y and t directions, for all temporal scales considered.

Similarly, in [52], underlying motion patterns are applied to identify video segments similar to a query sequence. This is done by computing the correlation of such motion patterns in the query video segment with a larger video sequence. The peaks in the correlation surface correspond to similar sequences. In this approach, the motion is estimated from the gradients inside small spatio-temporal patches or cuboids, instead of relying on expensive flow computations.

The proposal of [82] computes the similarity between images or videos by matching local self-similarities. Those are computed at pixel level, taking into account the similarity between a small patch around the considered pixel and a larger region surrounding it.

Also aimed at SFF applications, [83] propose to recognize actions from a unique example, by using local regression kernels based on weights computed on the video pixels and their neighbors both in space and time.

In [84], the optical flow computed for the entire video is represented by magnitude and orientation. A histogram is built on the quantized orientation, using only pixels for which the flow magnitude is above a certain threshold. Also, the flow of the considered pixels is weighted by their magnitude. The normalized histograms for the training sequences for each class are submitted to PCA for dimensionality reduction.

In [85], motion vectors from the compressed-domain are used to estimate motion fields, which are then submitted to a hierarchical agglomerative clustering algorithm, in order to create an organizing hierarchy for videos which are presumed to be based on actions.

In contrast to the BoVF-based approaches (see Section 5.2.1 below), [86] argue that human actions should be characterized by large-scale features instead of local patches. Therefore, the authors consider the frame as the basic

unit for initial description, which is made in terms of the motion descriptors proposed by Efros in [67]. Their “visual vocabulary” is then built on those global frame features, whose space is quantized by the k-medoid clustering algorithm. Finally, each video sequence is represented in terms of the frequency of such “frame-words”.

In [87], a hierarchical space-time model is implemented in two layers: the bottom layer of features composed of a bank of 3D Gabor filters; the second layer in the proposed hierarchy are histograms of Gabor orientations. This proposal is based on that of [114] for object recognition, which tries to mimic organic visual systems, which are seen as being composed of two kinds of brain cells with different roles in the recognition process.

Finally, in [88], the temporal evolution of Histograms of optical Flow (HoF) features gathered from each frame are modeled by a non-linear dynamical system.

5.2. Approaches Based on Statistics of Local Descriptors

Last section discussed the first attempts at avoiding the constraints imposed by object models for action recognition. However, most of those approaches either rely on domain knowledge or are focused on constrained settings or databases, like surveillance or Smart Fast-Forward (SFF) applications. Another drawback of those approaches is that, being based on dense low-level features, they demand great computational effort.

To mitigate those drawbacks, approaches based on mid-level local descriptors, mostly computed on a (potentially small) number of interest points emerged as a promising trend for action recognition. More specifically, approaches based on histograms of quantized local descriptors – known as Bag of Visual Features (BoVF)³ – have shown to provide consistently good results reported by a number of independent authors in a variety of scenarios, including datasets composed of professional and amateur realistic videos.

Despite the success of BoVF-based approaches, there are a few other strategies to gather information from local descriptors. These alternative strategies are gathered in a separate category in the proposed framework.

³Due to the lack of standard terminology, those approaches have also been denominated bag of visual words, bag of keypoints, bag of features or bag of words in the literature.

5.2.1. Using Bag of Visual Features (BoVF)

BoVF representations are inspired by traditional textual Information Retrieval (IR) techniques, in which the feature vectors that represent each text document in a collection are histograms of word occurrences [115]⁴. Such representation is referred to as Bag-of-Words (BoW), in order to emphasize that it is comprised of orderless features.

A remarkable difference in the analogy between BoVFs and BoWs is the need to define what constitutes a *visual word*. Such “definition” is achieved in practice by a process called *vocabulary* (or *codebook*) *learning*, consisting of the quantization of the descriptors’ feature space, typically computed by clustering. A detailed introduction on how BoVF representations are build both for images and videos can be found in [89].

BoVF-based approaches have been first applied to object classification and have proved very robust to background clutter, occlusion and scale changes, indicating their potential for challenging object recognition settings ([116], [117], [118], [119], [120], [105]). BoVF and its variations have demonstrated similar strengths when applied to action recognition, thus becoming, by far, the most common base representation found on recent proposals.

The relevance of BoVF-based action recognition is reinforced by the fact that those schemes became a common testbed for several spatio-temporal points detection and description algorithms. The work of [121], for example, compares different alternatives for interest point detectors/descriptors applied in a classic BoVF representation for Internet videos. Similar comparisons can also be found in [120], [122], [123], [124] and [125].

To the best of our knowledge, the approach proposed by [90] is the seminal work on BoVF techniques applied to action recognition. For the low-level features, the spatio-temporal interest points proposed in [126] are described by spatio-temporal jets. K-means clustering is applied to create the quantized vocabulary, based on which the histogram of local features is computed. This is also the work which introduced the Royal Institute of Technology (KTH) action database, which later became a *de facto* standard for action recognition algorithms. The work described in [90] is extended in [91], which proposes a mechanism for local velocity adaptation aimed at compensation of camera motion that could affect local measurements.

⁴In fact, each histogram bin reflects not a single word, but a family of words represented by their roots.

The work of [53] extends previous work on object recognition based on sparse sets of feature points. The interest points selection method applied in this work is based on separable linear filters. Three descriptors for the cuboids delimited around the selected points are tested: *normalized pixel values*, *brightness gradients* and a *windowed optical flow*. Principal Component Analysis (PCA) is used for dimensionality reduction of the point descriptors and a typical BoVF signature is then built on them. The k-means clustering algorithm is used for defining the dictionary.

Another BoVF approach is proposed by [92], this time based on an extension of the Scale-Invariant Feature Transform (SIFT) descriptor [127]. The new descriptor adds temporal information, extending the original SIFT descriptor to a 3D space. Instead of using the SIFT detector, points are selected at random. Histograms built on a codebook created with k-means are the initial signatures. Then, to create an enhanced representation, a criteria based on the co-occurrence of visual words is applied to reduce the vector dimension. In other words, those visual words which co-occur above a certain threshold are joined.

In [93] the local descriptors are based on the responses to a bank of 3D Gabor Filters, followed by a MAX-like operation. Such features are computed on patches delimited by a sliding window and described by histograms generated by the quantization of the orientations in nine directions. The quantization of those histograms into a codebook is learned from a gaussian mixture model.

In [94], a BoVF representation based on the features proposed by Dollar [53] is used together with generative models – unlike previously discussed methods which are based on discriminative ones – for action recognition. Two methods borrowed from traditional textual Information Retrieval research are examined: probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA).

In [95], a joint framework using BoVFs both for scene and actions recognition is proposed. The underlying assumption is that scenes can provide contextual information to improve the recognition of some actions classes. Initially, movie scripts provide for automatic annotation of scenes and actions. Text mining is then applied to discover co-occurrences between scenes and actions. Finally, separate Support Vector Machines (SVM) models based on BoVFs are learned using the same approach described in [96].

BoVF Variations

A number of variations over the typical BoVF scheme have been proposed, mostly aimed at dealing with specific recognized limitations of pure BoVF-based approaches, the main ones being the lack of structural information and the poor quality of the visual vocabulary.

In [45], the lack of structural information of BoVF representations is addressed in a rather direct manner: each frame is subdivided into cells, over which BoVF based on Dollar’s features [53] are computed. Additionally, motion features from neighbor frames are used in a weighted scheme which takes into account the distance of the neighbor from the actual frame. A spatial-pyramid matching, similar to the one used in [117], is then applied to compute the similarity between frames. Finally, frames are classified individually, and a voting scheme decides the final video classification.

In [96], it is proposed a BoVF representation built from the Space-Time Interest Points (STIP) presented in [126], but without scale selection. The points are described by Histograms of Gradients (HoG) and Histograms of optical Flow (HoF), computed over the spatio-temporal volumes positioned around them. To add some structural information, each video volume is subdivided by a grid, so that at recognition time different configurations for the grids are considered, using a multi-channel SVM classifier.

Most authors working on BoVF-based approaches for actions recognition learn the vocabulary by using the k-means clustering preceded by a PCA-based dimensionality reduction. Nevertheless, some alternatives to vocabulary learning have been proposed, either by enhancing the vocabulary delivered by k-means or by applying alternative clustering techniques.

The work of [97] merges k-means results to produce an enhanced vocabulary, indicating that better vocabularies can have a significant impact on recognition. Using features similar to those of Dollar [53], they propose a Maximization of Mutual Information (MMI) criteria to merge the cuboid clusters output by k-means. Those new clusters are called Video Words Clusters (VWC). Additionally, two approaches to add structural information are explored: spatial correlogram, with the distances quantized in a few levels and the Spatio-Temporal Pyramid Matching (STPM) of [117].

Another example of the impact of a better vocabulary is found in [128], where the authors proposed an enhanced BoVF in which the relationships among visual words are explored. This is pursued with the aid of a visual ontology inspired on WordNet [129], a textual ontology extensively applied

for text retrieval. Their visual ontology is built by applying agglomerative clustering to the visual words previously discovered by k-means. From that, it is possible to compute the specificity (i.e., the depth in the ontology tree), path length (i.e., the number of links in the path between two words) and information content (relative to the probability of word occurrences). Those precomputed values are used in a soft-weighting scheme aiming at better evaluating the significance of each word.

In [130], a BoVF combining dynamic and static local features is proposed to address action recognition in YouTube videos. Static information captured by three different interest point detectors are described by SIFT descriptors. Motion is collected by Dollar’s interest points [53], described by gradient vectors. The spatio-temporal distribution of motion features is used to localize coarse ROIs, which are used together with the PageRank algorithm, for pruning spurious features. In addition to such motion-guided feature selection, the authors propose a procedure to create a semantic visual vocabulary, which involves enhancing the result of k-means by using a technique based on the KL-divergence algorithm.

In [49], a modified version of k-means which takes into account the spatial localization of the interest points is applied to form the codebook. Additionally, a 3D extension of the Harris detector alternative to that of Laptev [126] is proposed, aimed at selecting a denser sampling. Cuboids defined around the detected points are described in terms of shape and motion.

Finally, in [98], the visual word distribution is described by Gaussian Mixture Models (GMM) of SIFT descriptors. These GMMs are specialized for each video clip, and a kernel for video comparison is built on the Kullback-Leibler divergence ([131] *apud* [98]).

5.2.2. *Alternative Strategies to Capture Information from Local Descriptors*

Despite the great success of BoVF-based approaches and their variations, achieving high recognition rates in truly realistic databases remains an open challenge. A number of authors have been trying alternative ways to collect relevant information out of local descriptors.

In [99], the salient-points detected at peaks of activity are clustered into salient regions, whose scales are correlated to the motion magnitude. Noisy interest points are discarded, so the videos are described by remaining points inside detected salient regions. Since such representations do not have the same dimension for all video sequences, the Chamfer distance is used as the kernel for a Relevance Vector Machines (RVM) classification algorithm. A

space-time warping adjustment scheme is applied to deal with varied execution speeds. Some variations to this overall scheme are presented in [100].

In [101], it is argued that correlograms can capture the spatial arrangement of codewords in the case of object classification. This is applied in [132], in which an extension of the spatial correlatons (quantized in *correlograms*) is proposed for action recognition. Rather than building a histogram of interest-points, as in typical BoVF approaches, action is modeled as a collection of space-time interest points where each interest point has a label from the vocabulary of video-words. So, the video sequences are composed of sets of video words and their spatio-temporal relations are described in form of spatio-temporal correlatons. Actions are modeled by estimating – with pLSA – the codewords distribution for each particular class.

In [102], it is observed that the lack of structural information also means the absence of temporal sequencing. The representation proposed uses Dollar’s features [53], but the histogram built on them is made up of temporal bins. PrefixSpan algorithm is used for mining frequent sequences and the LPBoost algorithm is used to identify the most discriminative ones.

In [55], dense corner features are hierarchically grouped in space and time to produce a compound feature set. Data mining is applied to group features in multiple stages, from the initial low-level features until a higher level in which the relative positions of groupings are used. As in [89], 2D features are collected from the planes defined by the coordinates (x, y) – the frames – (x, t) and (x, t) , which are, in the former case, considered as distinct channels. Motion is captured by dominant orientation and points are described only by their scale, corresponding channel and orientation, instead of more complex descriptors like SIFT or STIP. Transaction vectors are built based on neighbor interest points and the Apriori algorithm is applied to the transactions in order to find association rules between transactions vectors and actions.

In [103], a matching kernel function for comparing spatio-temporal relationships among interest points is proposed for detection and localization of multiple actions and interactions in unsegmented videos. First, interest points are detected and described as usual. Afterwards, pairwise relationship predicates are used to describe the structural relations. Temporal relations are described by Allen’s taxonomy (*equals*, *before*, *meets*, *overlaps*, *during*, *starts* and *finishes*) [133] *apud* [103], with respect to the interval limits given by the volume patch dimensions projected onto the temporal axis. Similar spatial predicates are created, so that temporal and spatial 3D relationship

histograms aimed at capturing both appearance and point relationships in the video can be computed. Finally, the proposed matching kernel captures the similarity between two histograms by their intersection.

The work presented in [104] proposes a video representation in the form of a vocabulary-tree, based on the outputs of several interest point detectors plus a dense sampling for action recognition and localization. Motion compensation is achieved by using previously tracked features to perform the segmentation of the image into motion planes. Such a segmentation is performed by an initial color-based segmentation followed by homography computation using RANSAC. The homography is then used to correct the motion of the features inside each dominant plane.

The fact that humans can recognize actions just by observing some tracked points has been explored in several approaches relying on trajectory models of points placed at specific body parts. The work of [105] extends this notion into a more generalizable approach, in which the authors propose to gather information about the spatio-temporal context of tracked SIFT points in a hierarchical three-level scheme. In the first level – the point-level context – local statistics of gradients along point trajectories are computed. In the second level – the intra-trajectory context – the dynamic aspects of those trajectories in the spatio-temporal space are considered. Finally, in the coarser level – the inter-trajectory context – the information about the spatio-temporal co-occurrences of trajectories distributions is collected.

In a similar vein, the tracks of a set of features – detected and tracked by the algorithm proposed in [134], but with weaker constraints – is employed by [25]. Such trajectories are described by the history of their quantized velocities.

5.3. *Using Concept Probabilities*

Using similar ideas from CBVR systems [2], some authors have proposed to use higher level concepts as the building blocks of video representations aimed at action recognition.

In [42], 39 semantic concept detectors from LSCOM-lite – SVM classifiers based on raw color and texture features [135] – are applied to video I-frames. Then, the trajectories of those concepts in the concept space are analyzed by Hidden-Markov Models (HMM), one for each concept axis. Their work reinforces the results of [5], providing additional evidence that, for some concepts, the dynamic information is essential. The authors found that dynamic infor-

mation enhanced recognition results of the following concepts: *riot*, *exiting car* and *helicopter hovering*.

In [106], Moving Picture Experts Group (MPEG) motion vectors are summarized in a motion image which describes the global motion pattern of a video shot. Motion images are combined with color and texture features and used as input for several weak SVM classifiers. The output of such classifiers are fused together to compose the video feature vector. The approach is tested on the TRECVID-2005 dataset, for selected dynamic concepts only, comparing favorably with results based on motion direction histograms and motion magnitude histograms.

In [107], 374 concepts are selected from the LSCOM ontology to be detected by three different SVM detectors based on histograms of low-level features (grid color moments, Gabor textures and edge direction histograms). The results of those classifiers are fused together in order to produce scores for each concept. Variations in the duration of action clips are dealt with by applying the Earth Mover’s Distance (EMD) in multiple temporal scales.

A framework for event detection presented in [46] starts by the application of BoVF-based approach to detect a number of concepts. Relative motion of keypoints between successive frames is used to aid the spatial clustering of visual words. A visual word ontology is then built based on the output of the spatial clustering, in order to take into account the correlation of visual words potentially related to the same object or object parts. The final representation is a collection of BoVFs built on those roughly segmented regions.

6. Hybrid Approaches

This section is devoted to action recognition approaches that fuse information coming from both object models and global statistical representations. It is worth noticing that, from the point of view of generalization ability, such mixed approaches are limited by the representation whose computation imposes stronger constraints.

In [108], it is proposed a hybrid approach where constellation models are used to add geometric information to the classical BoVF representation. This is done by modeling actions within a two-layered hierarchical model. The higher layer is comprised of selected body parts, which are then described as BoVFs. The BoVF-based system proposed is based on Dollar’s features [53], together with sampled edge points described by shape context.

The work presented in [109] mixes low-level, local descriptors and shape-based representations. They build several vocabulary trees on points selected by five different 2D point selectors. To include dynamic information, motion maps are obtained from a Lucas-Kanade optical flow computation, and motion is represented by velocity maps between pairs of frames. A technique for compensation of camera motion based on a global similarity transformation is also presented and applied. A star-shape model – aimed at coarsely capturing some structural information of the moving object – is used to guide the process.

In [110], the concept of Self-Similarity Matrix (SSM) is introduced to build video representations aimed at action recognition. An SSM is a table of distances between all video frames. Although this definition can be applied for any feature type, in [110], they are computed over trajectories of human joints, which are fused together with those computed from HoG and HoF features. The final descriptor is obtained by considering the SSM sequences as images and splitting them into patches, which are described by histograms of gradient directions.

The proposal of [111] works on clouds of interest points collected over different time scales. The distribution of such clouds in both space and time is described by global features. To compose the clouds, they propose a spatio-temporal interest point detector that collects dense samplings of interest points. In order to avoid too many spurious point detections, the moving object is coarsely separated from the background.

In [112], BoVFs of 2D and 3D SIFT feature descriptors, extracted on 2D SIFT interest points, as well as Zernike moments, are applied to both frames and MEI images. The extraction of those features are guided by frame subtraction, which provides a coarse motion-based segmentation. Feature fusion is achieved by simple concatenation of the resulting four descriptors.

The work of [113] proposes another hybrid approach for both recognition and detection of actions in unsegmented videos. Visual codebooks are class-specific and take co-occurrences of visual words pairs into account. The positions of these visual words in relation to the object center are used to model the actions, which therefore implies the need for object segmentation. In addition, spatio-temporal scale adjustments are done manually for training. Finally, a framework for voting over time, which is based on optical flow and appearance descriptors, is proposed for action segmentation.

7. Summary and Discussion

Figure 2 together with Tables 2, 3 and 4 summarize our survey on previous and state-of-the-art action recognition approaches. As already discussed throughout the text, approaches relying on object models to describe video content have the drawback of imposing a number of constraints on the action scenario. Such constraints are rarely met in feature movies or user-generated videos found in video sharing systems (e.g., YouTube), thus pushing the research in action recognition towards more general approaches. It is important to notice, though, that some approaches based on object models have proved successful in realistic but restrict application domains, like surveillance and HCI, for example. In fact, provided that their constraints can be guaranteed, those approaches should be considered as potential choices in those cases where real-time processing is a requirement. In particular, approaches based on implicit models built on 2D descriptors – for instance, like those based on MEIs and MHIs [50] – tend to be quite efficient. In addition, the advances in the state-of-the-art of pre-processing techniques like segmentation and tracking can turn some approaches based on object models better suited for realistic environments, possibly giving rise to new hybrid approaches.

Regarding approaches based on global statistics, a clear tendency towards the usage of BoVF representations emerges, specially in those attempts to deal with unconstrained video databases.

Nevertheless, in spite of their success, a number of limitations are yet to be overcome by BoVF-based systems in order to achieve solutions that are mature enough to be incorporated in real-world tools. One of those issues is the *ad-hoc* nature of the visual vocabulary learning process. Although some papers discussed (Section 5.2.1) indicate better alternatives to pure k-means, there is no principled methodology neither to build an optimal vocabulary, nor even to preemptively assess the vocabulary quality, given a specific database.

The relatively small number of proposals dealing with local descriptors in alternative ways (Section 5.2.2) prevents the anticipation of specific trends in this direction, but it is worth noticing that both BoVF and non-BoVF-based approaches have been moved from an initial preference for sparse set of interest points to a current trend towards denser sets, on the assumption that they are more appropriate for realistic scenarios. This premise is reinforced, for example, by the comparison among several 3D point detectors and descriptors performed by [61], which concluded that, except for the

KTH database (which has very little contextual information, given its neutral background), regularly spaced dense samplings of points perform better at action recognition than interest points.

Approaches based on concept probabilities are somewhat underexplored for action recognition, if one considers their widespread usage in CBVR systems. Reasons for this apparent lack of interest might be the scarcity of annotated training samples for each concept classifier, as well as the issues raised by the usage of meta-classification and classification fusion.

The scarcity of labeled data for action recognition itself is an important issue which needs to be tackled in order to allow more significant advances in the area. Some initiatives have generated *de facto* standard databases, like Weizmann [51] and KTH [90] controlled databases, and, recently, more realistic databases like the Hollywood Movies Dataset ([96], [95]) and the Action Dataset [130]. Such annotation efforts have been fundamental to the research advances in the last few years, but they suffer the limitation of being somewhat isolated efforts and therefore, necessarily limited in size and scope. The TRECVID benchmark, which is well-known for its collective annotation efforts at larger scales, has had an event detection/recognition track for a few years, but videos and ground-truth data are available for participants only.

Besides publishing their own databases, some authors propose alternative ways for dealing with the issue of lack of annotated data. In [136], it is proposed a semi-automatic annotation technique based on movies scripts and subtitles. In [137], the issue of collecting a large-enough amount of training data for high-level video retrieval is addressed by the usage of videos collected from YouTube filtered by pre-defined categories and tags.

Semi-supervised methods can also be used to produce larger amounts of annotated data from a small number of samples. In [138], a template is manually generated by cropping bounding boxes from one action example in a clean, controlled database and then applied to detect other instances of the same action in cluttered videos, using Dynamic Time Warping (DTW) to deal with variations in the duration of the action. Similarly, in [139], action prototypes are learned on a laboratory dataset and then applied to action classification in videos with dynamic backgrounds.

Although most current approaches for action recognition focus on enhancing recognition capability, it is worth mentioning that in order to scale up to the dimensions of the web or of any large database, efficiency needs to be taken into account. Some efforts in this direction are available in the literature. In the work of [139], for example, lookup tables of action prototypes are

used to speed-up action classification. In [104], [140] and [109], the features are quantized using vocabulary trees, which lead to more efficient matching when compared with traditional codebooks. In the realm of more typical BoVF schemes, compact vocabularies – like those in [97], [128] and [130] – can help to reduce the overall computational effort. In [75], [141] and [43], efficient boosting algorithms are applied for volumetric matching. In [109], the high computational complexity of the proposed approach – based on several dense interest point detectors – is explicitly pointed out, and led the authors to implement their recognition framework in a parallel architecture.

8. Concluding Remarks

This survey attempts to summarize the efforts of the academic community at the task of recognizing human actions from videos, with emphasis on recent approaches. It proposes a new organizing framework, based on the representations chosen, and therefore, on their underlying assumptions. This organization allows to categorize the newest approaches smoothly alongside the more traditional ones. It also allows to compare and contrast different methods based on their constraints, which, we hope, enables a principled selection of a method, given the application domain. We observe that there is a correlation between our classification criteria and the chronology of methods, indicating a trend toward progressively weakening the constraints imposed on video content.

The greater focus on BoVF-based approaches emerges naturally from their potential on the field, making it a promising direction to pursue in the search for effective solutions for recognizing human actions in scenarios of realistic videos. Many questions, however, remain open, and a better assessment of the capabilities of current methods in very challenging scenarios will depend on a collective effort of generating annotated data.

Acknowledgements

The authors thank Dr. Ivan Laptev for his comments on earlier versions of this manuscript, as well as the Brazilian funding agencies CAPES, CNPq and FAPEMIG.

References

- [1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, *Trans. Pattern Anal. Mach. Intell.* 22 (12) (2000) 1349–1380.
- [2] C. G. M. Snoek, M. Worring, Concept-based video retrieval, *Foundations and Trends in Information Retrieval* 2 (4) (2008) 215–322.
- [3] L. Kennedy, A. Hauptmann, LSCOM Lexicon Definitions and Annotations (Version 1.0), Tech. rep., Columbia University (March 2006).
- [4] P. Over, T. Ianeva, W. Kraaij, A. F. Smeaton, Trecvid 2005 - an overview, in: *In Proceedings of TRECVID '05*, 2005.
- [5] L. Kennedy, Revision of LSCOM Event/Activity Annotations, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Tech. rep., Columbia University (December 2006).
- [6] G. Lavee, L. Khan, B. Thuraisingham, A framework for a video analysis tool for suspicious event detection, *Multimedia Tools Appl.* 35 (1) (2007) 109–123.
- [7] A. Kale, A. Sundaresan, A. Rajagopalan, N. Cuntoor, A. Roy Chowdhury, V. Kruger, R. Chellappa, Identification of humans using gait, *Trans. Image Proces.* 13 (9) (2004) 1163–1173.
- [8] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, *Trans. Sys., Man and Cybern.* 34 (3) (2004) 334–352.
- [9] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, H. Sun, Structure analysis of soccer video with domain knowledge and hidden markov models, *Pattern Recogn. Lett.* 25 (7) (2004) 767–775.
- [10] G. Zhu, Q. Huang, C. Xu, Y. Rui, S. Jiang, W. Gao, H. Yao, Trajectory based event tactics analysis in broadcast sports video, in: *Proceedings of ACM MULTIMEDIA '07*, 2007, pp. 58–67.
- [11] M. M. Rahman, S. Ishikawa, Human motion recognition using an eigenspace, *Pattern Recogn. Lett.* 26 (6) (2005) 687–697.

- [12] X. Tong, L. Duan, C. Xu, Q. Tian, H. Lu, Local motion analysis and its application in video based swimming style recognition, in: Proceedings of ICPR '06, 2006, pp. 1258–1261.
- [13] R. Tan, J. W. Davis, Differential video coding of face and gesture events in presentation videos, *Comput. Vis. Image Underst.* 96 (2) (2004) 200–215.
- [14] H. Cooper, R. Bowden, Sign language recognition using boosted volumetric features, in: Proceedings of IAPR MVA '07, 2007, pp. 359–362.
- [15] P. Buehler, A. Zisserman, M. Everingham, Learning sign language by watching tv (using weakly aligned subtitles), in: Proceedings of IEEE CVPR '09, 2009, pp. 2961–2968.
- [16] S. Sundaram, W. Cuevas, High level activity recognition using low resolution wearable vision, in: Proceedings of IEEE CVPRW '09, 2009, pp. 25–32.
- [17] A. F. Bobick, S. S. Intille, J. W. Davis, F. Baird, C. S. Pinhanez, L. W. Campbell, Y. A. Ivanov, A. Schütte, A. Wilson, The kidsroom: A perceptually-based interactive and immersive story environment, *Presence: Teleoper. Virtual Environ.* 8 (4) (1999) 369–393.
- [18] H. Ren, G. Xu, Human action recognition in smart classroom, in: Proceedings of IEEE AFGR '02, 2002, pp. 399–404.
- [19] F. Tsalakanidou, S. Malassiotis, Robust facial action recognition from real-time 3d streams, *Proceedings of IEEE CVPRW '09* 0 (2009) 4–11.
- [20] J. Cohn, Use of active appearance models for analysis and synthesis of naturally occurring behavior, in: Proceedings of IEEE CVPRW '09, Vol. 0, 2009, pp. 1–3.
- [21] L. Zelnik-Manor, Statistical analysis of dynamic actions, *Trans. Pattern Anal. Mach. Intell.* 28 (9) (2006) 1530–1535, member-Michal Irani.
- [22] V. Kruger, D. Kragic, A. Ude, C. Geib, The meaning of action: a review on action recognition and mapping, *Advanced Robotics* 21 (September 2007) 1473–1501(29).

- [23] A. Branzan Albu, T. Beugeling, N. Virji Babul, C. Beach, Analysis of irregularities in human actions with volumetric motion history images, in: Motion '07, 2007, p. 16.
- [24] G. Kosta, M. Benoit, Group behavior recognition for gesture analysis, Trans. Circ. Syst. Video Techn. 18 (2) (2008) 211–222.
- [25] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: Proceedings of ICCV '09, 2009.
- [26] C. Cedras, M. Shah, Motion-based recognition: A survey, Image and Vision Computing 13 (2) (1995) 129–155.
- [27] J. Aggarwal, Q. Cai, Human motion analysis: a review, in: Proceedings of IEEE Nonrigid and Articulated Motion Workshop '97, 1997, pp. 90–102.
- [28] V. I. Pavlovic, R. Sharma, T. S. Huang, Visual interpretation of hand gestures for human-computer interaction: A review, Trans. Pattern Anal. Mach. Intell. 19 (1997) 677–695.
- [29] D. M. Gavrilu, The visual analysis of human movement: a survey, Comput. Vis. Image Underst. 73 (1) (1999) 82–98.
- [30] M. Shah, Understanding human behavior from motion imagery, Mach. Vision Appl. 14 (4) (2003) 210–214.
- [31] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, Pattern Recog. 36 (3) (2003) 585–601.
- [32] Learning and understanding dynamic scene activity: a review, Image and Vision Computing 21 (1) (2003) 125 – 136.
- [33] J. Wang, S. Singh, Video analysis of human dynamics: A survey, Real-Time Imaging 9 (5) (2003) 320–345.
- [34] J. K. Aggarwal, S. Park, Human motion: Modeling and recognition of actions and interactions, in: Proceedings of 3DPVT '04, 2004, pp. 640–647.

- [35] R. Chellappa, A. K. Roy-Chowdhury, S. K. Zhou, Recognition of humans and their activities using video, *Synth. Lect. on Img, Vid. Multim. Process.* 1 (1) (2005) 1–173.
- [36] T. B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, *Comput. Vis. Image Underst.* 104 (2) (2006) 90–126.
- [37] R. Poppe, Vision-based human motion analysis: An overview, *Comput. Vis. Image Underst.* 108 (1-2) (2007) 4–18.
- [38] M. A. R. Ahad, J. Tan, H. Kim, S. Ishikawa, Human activity recognition: Various paradigms, in: *Proceedings of ICCAS '08*, 2008, pp. 1896–1901.
- [39] P. Turaga, R. Chellappa, V. Subrahmanian, O. Udrea, Machine recognitionnology of human activities: A survey, *Trans. Circuits and Sys. for Video Tech.* 18 (11) (2008) 1473–1488.
- [40] W. Ren, S. Singh, M. Singh, Y. S. Zhu, State-of-the-art on spatio-temporal information-based video retrieval, *Pattern Recogn.* 42 (2) (2009) 267–282.
- [41] R. Poppe, A survey on vision-based human action recognition, *Image Vision Comput.* 28 (6) (2010) 976–990.
- [42] S. Ebadollahi, L. Xie, S. fu Chang, J. Smith, Visual event detection using multi-dimensional concept dynamics, in: *Proceedings of IEEE ICME '06*, Vol. 0, 2006, pp. 881–884.
- [43] I. Laptev, P. Perez, Retrieving actions in movies, in: *Proceedings of ICCV '07*, 2007, pp. 1–8.
- [44] Y. Ke, R. Sukthankar, M. Hebert, Event detection in crowded videos, in: *Proceedings of IEEE ICCV '07*, 2007, pp. 1–8.
- [45] Z. Zhao, A. Elgammal, Human activity recognition from frame’s spatiotemporal representation, in: *Proceedings of IEEE ICPR '08*, 2008, pp. 1–4.

- [46] F. Wang, Y.-G. Jiang, C.-W. Ngo, Video event detection using motion relativity and visual relatedness, in: Proceedings of ACM MULTIMEDIA '08, 2008, pp. 239–248.
- [47] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, M. de Miranda Coelho, A. de Albuquerque Araújo, Nude detection in video using bag-of-visual-features.
- [48] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, in: Proceedings of IEEE CVPR '05, Vol. 1, 2005, pp. 984–989.
- [49] D. Chen, H. Wactlar, M.-y. Chen, C. Gao, A. Bharucha, A. Hauptmann, Recognition of aggressive human behavior using binary local motion descriptors, in: Proceedings of IEEE EMBS '08, 2008, pp. 5238–5241.
- [50] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, *Trans. Pattern Anal. Mach. Intell.* 23 (3) (2001) 257–267.
- [51] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Proceedings of IEEE ICCV '05, Vol. II, 2005, pp. 1395–1402.
- [52] E. Shechtman, M. Irani, Space-time behavior based correlation, in: Proceedings of IEEE CVPR '05, 2005, pp. 405–412.
- [53] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of ICCCN '05, 2005, pp. 65–72.
- [54] S. Ebadollahi, L. Xie, S. fu Chang, J. Smith, Visual event detection using multi-dimensional concept dynamics, in: Proceedings of IEEE ICME '06, 2006, pp. 881–884.
- [55] A. Gilbert, J. Illingworth, R. Bowden, Scale invariant action recognition using compound features mined from dense spatio-temporal corners, in: Proceedings of ECCV '08, 2008, pp. 222–233.
- [56] P. Peursum, H. H. Bui, S. Venkatesh, G. West, Human action segmentation via controlled use of missing data in hmms, in: Proceedings of IEEE ICPR '04.

- [57] P. Peursum, H. H. Bui, S. Venkatesh, G. West, Robust recognition and segmentation of human actions using hmms with missing observations, *J. Appl. Signal Process.* 2005 (1) (2005) 2110–2126.
- [58] H. Jiang, M. Drew, Z. Li, Successive convex matching for action detection, in: *Proceedings of IEEE CVPR '06*, Vol. II, 2006, pp. 1646–1653.
- [59] G. Kosta, C. Pedro, M. Benoit, Modelization of limb coordination for human action analysis, in: *Proceedings of IEEE ICIP '06*, 2006, pp. 1765–1768.
- [60] R. Filipovych, E. Ribeiro, Learning human motion models from unsegmented videos, in: *Proceedings of IEEE CVPR '08*, 2008, pp. 1–7.
- [61] H. Wang, M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: *Proceedings of BMVA BMVC '09*, 2009, pp. 1–5.
- [62] P. Dhillon, S. Nowozin, C. Lampert, Combining appearance and motion for human action classification in videos, in: *Proceedings of IEEE CVPRW '09*, Vol. 0, 2009, pp. 22–29.
- [63] M. F. Abdelkader, A. K. Roy-Chowdhury, R. Chellappa, U. Akdemir, Activity representation using 3d shape models, *J. Image Video Process.* 8 (2) (2008) 1–16.
- [64] N. Cuntoor, B. Yegnanarayana, R. Chellappa, Activity modeling using event probability sequences, *Trans. Image Process.* 17 (4) (2008) 594–607.
- [65] M. Hu, S. Ali, M. Shah, Detecting global motion patterns in complex videos, in: *Proceedings of IEEE ICPR '08*, 2008, pp. 1–5.
- [66] M. Hu, S. Ali, M. Shah, Learning motion patterns in crowded scenes using motion flow field, in: *Proceedings of IEEE ICPR '08*, 2008, pp. 1–5.
- [67] A. A. Efros, A. C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: *Proceedings of ICCV '03*, 2003, pp. 726–733.

- [68] K. Hatun, P. Duygulu, Pose sentences: A new representation for action recognition using sequence of pose words, in: Proceedings of IEEE ICPR '08, 2008, pp. 1–4.
- [69] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, T. Huang, Action detection in complex scenes with spatial and temporal ambiguities, in: Proceedings of IEEE ICCV '09, 2009.
- [70] S.-F. Wong, T.-K. Kim, R. Cipolla, Learning motion categories using both semantic and structural information, in: Proceedings of IEEE CVPR '07.
- [71] C. Thureau, V. Hlavac, Pose primitive based human action recognition in videos or still images, in: Proceedings of IEEE CVPR '08, 2008, pp. 1–8.
- [72] A. Mokhber, C. Achard, M. Milgram, Recognition of human behavior by space-time silhouette characterization, *Pattern Recogn. Lett.* 29 (1) (2008) 81–89.
- [73] N. Cuntoor, Morse functions for activity classification using spatiotemporal volumes, in: Proceedings of CVPRW '06, 2006, p. 20.
- [74] A. Fathi, G. Mori, Action recognition by learning mid-level motion features, in: Proceedings of IEEE CVPR '08, 2008, pp. 1–8.
- [75] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: Proceedings of IEEE ICCV '05, 2005, pp. 166–173.
- [76] G. Papadopoulos, V. Mezaris, I. Kompatsiaris, M. Strintzis, Estimation and representation of accumulated motion characteristics for semantic event detection, in: Proceedings of IEEE ICIP '08, 2008, pp. 41–44.
- [77] M.-C. Tien, Y.-T. Wang, C.-W. Chou, K.-Y. Hsieh, W.-T. Chu, J.-L. Wu, Event detection in tennis matches based on video data mining, in: Proceedings of IEEE ICME '08, 2008, pp. 1477–1480.
- [78] C. G. M. Snoek, M. Worring, Multimodal video indexing: A review of the state-of-the-art, *Multimedia Tools Appl.* 25 (1) (2005) 5–35.

- [79] E. Ermis, V. Saligrama, P. Jodoin, J. Konrad, Motion segmentation and abnormal behavior detection via behavior clustering, in: *Proceedings of IEEE ICIP '08*, 2008, pp. 769–772.
- [80] Y. Ma, P. Cisar, Event detection using local binary pattern based dynamic textures, Vol. 0, 2009, pp. 38–44.
- [81] L. Zelnik-Manor, M. Irani, Event-based analysis of video, in: *Proceedings of IEEE CVPR '01*, Vol. 2, 2001, pp. 123–130.
- [82] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: *Proceedings of IEEE CVPR '07*, 2007, pp. 1–8.
- [83] H. J. Seo, P. Milanfar, Detection of human actions from a single example, in: *Proceedings of IEEE CVPR '09*, 2009, pp. 1965–1970.
- [84] X. Li, Hmm based action recognition using oriented histograms of optical flow field, *Elect. Lett.* 43 (10) (2007) 560–561.
- [85] P. Ahammad, C. Yeo, K. Ramchandran, S. Sastry, Unsupervised discovery of action hierarchies in large collections of activity videos, in: *Proceedings of IEEE MMSP '07*, 2007, pp. 410–413.
- [86] Y. Wang, P. Sabzmeydani, G. Mori, Semi-latent dirichlet allocation: A hierarchical model for human action recognition, in: *Proceedings of HUMO '07*, 2007, pp. 240–254.
- [87] H. Ning, T. X. Han, D. B. Walther, M. Liu, T. S. Huang, Hierarchical space-time model enabling efficient search for human actions, *Trans. Cir. and Sys. for Video Technol.* 19 (6) (2009) 808–820.
- [88] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: *Proceedings of IEEE CVPR '09*, 2009, pp. 1932–1939.
- [89] A. P. B. Lopes, R. S. Oliveira, J. M. de Almeida, A. de Albuquerque Araújo, Comparing alternatives for capturing dynamic information in bag of visual features approaches applied to human actions recognition, in: *Proceedings of IEEE MMSP '09*, 2009.

- [90] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of IEEE ICPR '04, Vol. III, 2004, pp. 32–36.
- [91] I. Laptev, B. Caputo, C. Schuldt, T. Lindeberg, Local velocity-adapted motion events for spatio-temporal recognition, *Comput. Vis. Image Underst.* 108 (3) (2007) 207–229.
- [92] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of ACM MULTIMEDIA '07, 2007, pp. 357–360.
- [93] H. Ning, Y. Hu, T. Huang, Searching human behaviors using spatial-temporal words, in: Proceedings of IEEE ICIP '07, 2007, pp. 337–340.
- [94] J. C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, *Int. Journal Comp. Vision* 79 (3) (2008) 299–318.
- [95] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: Proceedings of IEEE CVPR '09, 2009, pp. 2929–2936.
- [96] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of IEEE CVPR '08, 2008, pp. 1–8.
- [97] J. Liu, M. Shah, Learning human actions via information maximization, in: Proceedings of IEEE CVPR '08, 2008.
- [98] X. Zhou, X. Zhuang, S. Yan, S.-F. Chang, M. Hasegawa-Johnson, T. S. Huang, Sift-bag kernel for video event analysis, in: Proceedings of ACM MULTIMEDIA '08, 2008, pp. 229–238.
- [99] A. Oikonomopoulos, I. Patras, M. Pantic, Spatiotemporal saliency for human action recognition, in: Proceedings of IEEE ICME '05, 2005, pp. 1–4.
- [100] A. Oikonomopoulos, I. Patras, M. Pantic, Kernel-based recognition of human actions using spatiotemporal salient points, in: Proceedings of IEEE CVPR-V4HCI '06, 2006, p. 151.

- [101] S. Savarese, J. Winn, A. Criminisi, Discriminative object class models of appearance and shape by correlators, in: Proceedings of IEEE CVPR '06, 2006, pp. 2033–2040.
- [102] S. Nowozin, G. Bakir, K. Tsuda, Discriminative subsequence mining for action classification, in: Proceedings of IEEE ICCV '07, 2007, pp. 1–8.
- [103] M. S. Ryoo, J. K. Aggarwal, Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, in: Proceedings of ICCV '09, 2009.
- [104] H. Uemura, S. Ishikawa, K. Mikolajczyk, Feature tracking and motion compensation for action recognition, in: Proceedings of BMVA BMVC '08, 2008, pp. 1–8.
- [105] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, in: Proceedings of IEEE CVPR '09, 2009, pp. 2004–2011.
- [106] A. Haubold, M. Naphade, Classification of video events using 4-dimensional time-compressed motion features, in: Proceedings of ACM CIVR '07, 2007, pp. 178–185.
- [107] D. Xu, S.-F. Chang, Video event recognition using kernel methods with multilevel temporal alignment, *Trans. Pattern Anal. Mach. Intell.* 30 (11) (2008) 1985–1997.
- [108] J. Niebles, F. Li, A hierarchical model of shape and appearance for human action classification, in: Proceedings of IEEE CVPR '07, 2007, pp. 1–8.
- [109] K. Mikolajczyk, H. Uemura, Action recognition with motion-appearance vocabulary forest, in: Proceedings of IEEE CVPR '08, 2008, pp. 1–8.
- [110] I. Junejo, E. Dexter, I. Laptev, P. Perez, Cross-view action recognition from temporal self-similarities, in: Proceedings of ECCV '08, Vol. II, 2008, pp. 293–306.

- [111] M. Bregonzio, S. Gong, T. Xiang, Recognising action as clouds of space-time interest points, in: Proceedings of IEEE CVPR '09, 2009, pp. 1948–1955.
- [112] X. Sun, M. Chen, A. Hauptmann, Action recognition via local descriptors and holistic features, in: Proceedings of IEEE CVPR-4HB '09, 2009, pp. 58–65.
- [113] A. Oikonomopoulos, I. Patras, M. Pantic, An implicit spatiotemporal shape model for human activity localization and recognition, in: Proceedings of IEEE CVPR-4HB '09, 2009, pp. 27–33.
- [114] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms, *Trans. Pattern Anal. Mach. Intell.* 29 (3) (2007) 411–426.
- [115] R. A. Baeza-Yates, B. A. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press / Addison-Wesley, 1999.
- [116] S. Agarwal, A. Awan, Learning to detect objects in images via a sparse, part-based representation, *Trans. Pattern Anal. Mach. Intell.* 26 (11) (2004) 1475–1490, member-Dan Roth.
- [117] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Proceedings of IEEE CVPR '06, Vol. II, 2006, pp. 2169–2178.
- [118] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, *Int. J. Comput. Vision* 73 (2) (2007) 213–238.
- [119] Y.-G. Jiang, C.-W. Ngo, J. Yang, Towards optimal bag-of-features for object categorization and semantic video retrieval, in: Proceedings of ACM CIVR '07, 2007, pp. 494–501.
- [120] S.-F. Wong, R. Cipolla, Extracting spatiotemporal interest points using global information, in: Proceedings of IEEE ICCV '07, 2007, pp. 1–8.
- [121] G. Schindler, L. Zitnick, M. Brown, Internet video category recognition, in: Proceedings of InterNet '08, 2008, pp. 1–7.

- [122] A. Klaser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: Proceedings of BMVA BMVC '08, 2008.
- [123] G. Willems, T. Tuytelaars, L. Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Proceedings of ECCV '08, 2008, pp. 650–663.
- [124] B. Kaiser, G. Heidemann, Qualitative analysis of spatio-temporal event detectors, in: Proceedings of IEEE ICPR '08, 2008, pp. 1–4.
- [125] K. Rapantzikos, Y. Avrithis, S. Kollias, Dense saliency-based spatiotemporal feature points for action recognition, in: Proceedings of IEEE CVPR '09, 2009, pp. 1454–1461.
- [126] I. Laptev, T. Lindeberg, Space-time interest points, in: Proceedings of ICCV '03, 2003, pp. 432–439.
- [127] D. Lowe, Object recognition from local scale-invariant features, in: Proceedings of ICCV '99, 1999, pp. 1150–1157.
- [128] Y.-G. Jiang, C.-W. Ngo, Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval, *Comput. Vis. Image Underst.* 113 (3) (2009) 405–414.
- [129] C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, The MIT Press, 1998.
- [130] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos 'in the wild', in: Proceedings of IEEE CVPR '09, 2009, pp. 1996–2003.
- [131] P. J. Moreno, P. P. Ho, N. Vasconcelos, A kullback-leibler divergence based kernel for svm classification in multimedia applications, in: Proceedings of NIPS '03, Vancouver, Canada, 2003.
- [132] S. Savarese, A. DelPozo, J. Niebles, L. Fei-Fei, Spatial-temporal correlations for unsupervised action classification, in: Proceedings of IEEE WMVC '08, 2008, pp. 1–8.
- [133] J. F. Allen, G. Ferguson, *Actions and events in interval temporal logic*, Tech. rep., Rochester, NY, USA (1994).

- [134] B. D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of IJCAI '81, 1981, pp. 674–679.
- [135] M. R. Naphade, L. Kennedy, J. Kender, S. F. Chang, J. R. Smith, P. Over, A. Hauptmann, A light-scale concept ontology for multimedia understanding for trecvid 2005, Tech. rep., IBM Research (2005).
- [136] O. Duchenne, I. Laptev, J. Sivic, F. Bach, J. Ponce, Automatic annotation of human actions in video, in: ICCV '09, 2009.
- [137] A. Ulges, C. Schulze, M. Koch, T. M. Breuel, Learning automatic concept detectors from online video, Comput. Vis. Image Underst. In Press, Corrected Proof (2009) –.
- [138] B. Yao, S.-C. Zhu, Learning deformable action templates from cluttered videos, in: Proceedings of ICCV '09, 2009.
- [139] Z. Lin, Z. Jiang, L. S. Davis, Recognizing actions by shape-motion prototype trees, in: Proceedings of ICCV '09, 2009.
- [140] R. Ji, X. Sun, H. Yao, P. Xu, T. Liu, X. Liu, Attention-driven action retrieval with dtw-based 3d descriptor matching, in: Proceedings of ACM MULTIMEDIA '08, 2008, pp. 619–622.
- [141] Y. Ke, R. Sukthankar, M. Hebert, Spatio-temporal shape and flow correlation for action recognition, Proceedings of IEEE CVPR '07 (2007) 1–8.